RC24237 (W0704-059) April 12, 2007
Mathematics

IBM Research Report

On the Behrens-Fisher Problem: A Globally Convergent Algorithm and a Finite-Sample Study of the Wald, LR, and LM Tests

Alexandre Belloni
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Gustavo Didier
University of North Carolina
Chapel Hill, NC

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

ON THE BEHRENS-FISHER PROBLEM: A GLOBALLY CONVERGENT ALGORITHM AND A FINITE-SAMPLE STUDY OF THE WALD, LR AND LM TESTS

By Alexandre Belloni*

IBM T.J. Watson Research Center and MIT

AND

By Gustavo Didier

University of North Carolina-Chapel Hill

In this paper we provide a provably convergent algorithm for the multivariate Gaussian Maximum Likelihood version of the Behrens-Fisher problem. Our work builds upon a formulation of the log-likelihood function proposed by Buot and Richards [5]. Instead of focusing on the first order optimality conditions, the algorithm aims directly for the maximization of the log-likelihood function itself to achieve a global solution. Convergence proof and complexity estimates are provided for the algorithm. Computational experiments illustrate the applicability of such methods to high-dimensional data. We also discuss how to extend the proposed methodology to a broader class of problems.

We establish a systematic algebraic relation between the Wald, Likelihood Ratio and Lagrangian Multiplier Test ($W \geq LR \geq LM$) in the context of the Behrens-Fisher problem. Moreover, we use our algorithm to computationally investigate the finite-sample size and power of the Wald, Likelihood Ratio and Lagrange Multiplier Tests, which previously were only available through asymptotic results. The methods developed here are applicable to much higher dimensional settings than the ones available in the literature. This allows us to better capture the role of high dimensionality on the actual size and power of the tests for finite samples.

1. Introduction. The so-called Behrens-Fisher problem may be straightforwardly stated as follows.

> Given two independent random samples $X_1, ..., X_{N_1}$ and $Y_1, ..., Y_{N_2}$, test whether their respective population means μ_1 and μ_2 coincide in the case where their

*Research support from the IBM Herman Goldstein Fellowship is gratefully acknowledged.

AMS 2000 subject classifications: Primary 62H15

Keywords and phrases: Behrens-Fisher problem, high-dimensional data, hypothesis testing, algorithm, Wald Test, Likelihood Ratio Test, Lagrange Multiplier Test, size, power

covariances Σ_1 and Σ_2 are unknown.

Despite the deceiving simplicity of its form, this problem has motivated a wealth of literature that began with the original works of Behrens [1] and Fisher [8, 9], and includes Welch [27, 28], Scheffé [22, 23], Yao [30], Robbins, Simons and Starr [19], Subrahmanian and Subrahmanian [26], Cox [7], to name a few. For a review of the solutions for the BFP, see for instance Stuart and Ord [25] and Kim and Cohen [14]. The proposed solutions involve a myriad of different approaches, ranging from fiducial inference to Bayesian techniques.

In this paper, we are interested in the multivariate version of the Behrens-Fisher problem under Normality. In other words, X_i, Y_j above should be interpreted as d-dimensional Gaussian random vectors with (vector) means μ_1 and μ_2, and Σ_1, Σ_2 as their respective $d \times d$ covariance matrices. In this context, the Likelihood Ratio Test is a natural choice in face of the well-known asymptotic behavior of the test statistic. It turns out, though, that the maximization of the log-likelihood function without restrictive assumptions on the covariances (e.g., $\Sigma_1 = \Sigma_2$) is a non-trivial matter. In general, explicit solutions to the maximization procedure do not exist, and due to non-concavities in the objective function the solution to the system of first order likelihood equations can lead to local optima, as shown in Buot and Richards [5]. Numerical algorithms are available in the literature (see, for instance, Mardia, Kent and Bibby [17], and Buot and Richards [5]), but their convergence properties are unknown.

The purpose of this paper is two-fold. First, to propose a provably convergent algorithm, called Cutting Lines Algorithm (CLA), for the Gaussian Maximum Likelihood Behrens-Fisher Problem (BFP, for short). Second, to use the algorithm to investigate the finite sample properties - size and power - of the Likelihood Ratio Test and of the asymptotically equivalent Wald and Lagrange Multiplier Tests in the context of the BFP. Such properties are generally unknown, especially in high-dimensional contexts.

The CLA avoids the trap of local maxima, which haunts most approaches in the literature, by aiming directly for the maximization of the log-likelihood function itself. For this purpose, we make use of the expression for the log-likelihood function recently proposed by Buot and Richards [5], which is particularly suitable for numerical methods.

The general maximization strategy may be schematically characterized as follows.

(i) Lift the log-likelihood maximization problem into a higher-dimensional

setting by adding artificial variables and constraints. This new problem, the Lifted BFP, has the same solution as the original BFP;

(ii) Create a family of *convex* modifications (subproblems) of the Lifted BFP which we call Ellipsoidal Mean Estimation Problems (EMEP);

(iii) Solve a sequence of EMEP whose solutions (estimators of the mean) converge to the global solution of the Lifted BFP, i.e., the proper maximum likelihood estimator of the mean.

Step (i) is a common procedure in Continuous Optimization when one wishes to find a simpler (but equivalent) description for the problem in a higher-dimensional setting.

Step (ii) generates a family of *convex* problems which is computationally tractable (in particular, first order conditions are not only necessary but also sufficient). In fact, due to the particular structure of the EMEP, we are able to propose a specialized method which solves each problem in this family very efficiently both theoretically and in (computational) practice.

Step (iii) plays the crucial role of avoiding local maxima to ensure the global optimality. To achieve that, the algorithm relies on the particular geometry of the non-convexities associated with the problem. Such geometry allows for the construction of a sequence of approximations (based on supporting lines) to the log-likelihood function itself which can be efficiently optimized. We prove that the proposed method converges to a global solution. Furthermore, a simulation study provides strong numerical evidence of the suitability of the CLA for solving high-dimensional problems. Problems with dimension up to 1000 were solved in a couple of minutes.

We are particularly interested in the finite-sample properties of the Wald, Likelihood Ratio and Lagrange Multiplier Tests. We show that their respective test statistics satisfy systematic algebraic inequalities in the context of the BFP (such result is known for classical linear models; see Savin [21], Berndt and Savin [2], and Breusch [4]). However, the CLA makes it possible to go one step further and provide a Monte Carlo study of the actual size and the power of such tests. Our results illustrate that the Wald Test is the most sensitive among the three to the impact of dimensionality, followed by the Likelihood Ratio Test. Especially when the sample size is (relatively) small with respect to the dimension, the Wald and the Likelihood Ratio Tests tend to over-reject the null hypothesis when we use the χ^2 quantiles given by Wilks' Theorem. In contrast, the observed size of the Lagrange Multiplier Test seems to be rather robust with respect to dimensionality, with a slight tendency to under-reject the null hypothesis. Perhaps not surprisingly, these properties carry over to the power of the tests: for fixed sample sizes, the Wald Test displays higher power than the Likelihood Ratio Test, which

in turn displays higher power than the Lagrange Multiplier Test. However, the similar shapes of the observed power curves of the three tests seem to suggest that, with appropriate test size adjustment, the three tests may end up showing similar power properties. We also applied the Bartlett correction to the Likelihood Ratio Test as proposed by Yanagihara and Yuan [29]. The corrected test tends to under-reject the null-hypothesis, especially for high-dimensional data. Accordingly, it usually displays lower power than the Lagrange Multiplier Test.

The paper is organized as follows. Section 2 revisits the reformulation of the log-likelihood function proposed by Buot and Richards [5]. Section 3 recasts the log-likelihood maximization problem as a non-convex programming problem, and introduces the EMEP. Section 4 studies the geometry of the non-convexities associated with the log-likelihood function. Section 5 presents the CLA and its convergence analysis. Section 6 studies the finite-sample properties of the Wald, Likelihood Ratio, Lagrange Multiplier and the Bartlett-corrected Likelihood Ratio Tests. The Appendix contains: the pertinent Convex Analysis definitions; an explanation of the relation between the EMEP and the BFP; a special-purpose algorithm for solving the EMEP; and an alternative convergent algorithm, called Discretization Algorithm, for solving the BFP.

2. Reformulation of the Likelihood Function. Recall that our goal is to maximize the log-likelihood function of two independent random samples $\{X_i\}_{i=1}^{N_1}$ and $\{Y_i\}_{i=1}^{N_2}$, where $X_i \sim N(\mu, \Sigma_1)$ and $Y_j \sim N(\mu, \Sigma_2)$ are d-dimensional (random) vectors. From now on we assume that the sample covariance matrices S_1 and S_2 are invertible. The maximization problem means that we should find μ, Σ_1, and Σ_2 that maximize

$$
\begin{aligned}
l(\mu, \Sigma_1, \Sigma_2) \quad = \quad & -\frac{1}{2}\sum_{i=1}^{N_1}(X_i-\mu)'\Sigma_1^{-1}(X_i-\mu) - \frac{N_1}{2}\log\det\Sigma_1 - \\
& -\frac{1}{2}\sum_{i=1}^{N_2}(Y_i-\mu)'\Sigma_2^{-1}(Y_i-\mu) - \frac{N_2}{2}\log\det\Sigma_2,
\end{aligned} \tag{1}
$$

which is a highly non-linear function of μ, Σ_1, and Σ_2.

Recently, a more (computationally) tractable reformulation of (1) was proposed by Buot and Richards [5]. For the reader's convenience, we reproduce it here. Denote the vector sample means by

$$
\bar{X} = \frac{1}{N_1}\sum_{i=1}^{N_1} X_i \quad \text{and} \quad \bar{Y} = \frac{1}{N_2}\sum_{i=1}^{N_2} Y_i, \tag{2}
$$

and the sample covariance matrices by

$$S_1 = \frac{1}{N_1}\sum_{i=1}^{N_1}(X_i - \bar{X})(X_i - \bar{X})' \quad \text{and} \quad S_2 = \frac{1}{N_2}\sum_{i=1}^{N_2}(Y_i - \bar{Y})(Y_i - \bar{Y})'. \tag{3}$$

Let $\widehat{\mu}$ be some possible value, or estimator, of μ, and *define* the matrices, or covariance estimators, $\widehat{\Sigma}_1$ and $\widehat{\Sigma}_2$ as

$$\widehat{\Sigma}_1 = \frac{1}{N_1}\sum_{i=1}^{N_1}(X_i - \widehat{\mu})(X_i - \widehat{\mu})' \quad \text{and} \quad \widehat{\Sigma}_2 = \frac{1}{N_2}\sum_{i=1}^{N_2}(Y_i - \widehat{\mu})(Y_i - \widehat{\mu})' \tag{4}$$

(note that if $\widehat{\mu}^*$, $\widehat{\Sigma}_1^*$ and $\widehat{\Sigma}_2^*$ are the maxima of l, or equivalently, the maximum likelihood estimators of μ, Σ_1 and Σ_2, then relations (4) hold as a result, not just as a definition).

Through simple manipulation of the expressions for $\widehat{\Sigma}_1$ and $\widehat{\Sigma}_2$, we get

$$\widehat{\Sigma}_1 = S_1 + (\bar{X} - \widehat{\mu})(\bar{X} - \widehat{\mu})' \quad \text{and} \quad \widehat{\Sigma}_2 = S_2 + (\bar{X} - \widehat{\mu})(\bar{X} - \widehat{\mu})'. \tag{5}$$

This implies that, once the estimator $\widehat{\mu}$ is obtained, $\widehat{\Sigma}_1$ and $\widehat{\Sigma}_2$ are easily computed by a rank-one update of the sample covariance matrices.

Note that, for some positive definite matrix M and vector v,

$$\det(M+vv') = \det(M^{1/2}(I+M^{-1/2}vv'M^{-1/2})M^{1/2}) = \det(M)(1+v'M^{-1}v). \tag{6}$$

Thus, by using formula (6), we obtain

$$\begin{aligned} \det(\widehat{\Sigma}_1) &= \det(S_1)\left(1 + (\bar{X} - \widehat{\mu})' S_1^{-1}(\bar{X} - \widehat{\mu})\right) \quad \text{and} \\ \det(\widehat{\Sigma}_2) &= \det(S_2)\left(1 + (\bar{Y} - \widehat{\mu})' S_2^{-1}(\bar{Y} - \widehat{\mu})\right). \end{aligned} \tag{7}$$

Moreover, we have

$$\begin{aligned} \sum_{i=1}^{N_1}(X_i - \widehat{\mu})'\widehat{\Sigma}_1^{-1}(X_i - \widehat{\mu}) &= \sum_{i=1}^{N_1}\widehat{\Sigma}_1^{-1} \circ (X_i - \widehat{\mu})(X_i - \widehat{\mu})' \\ &= \widehat{\Sigma}_1^{-1} \circ \sum_{i=1}^{N_1}(X_i - \widehat{\mu})(X_i - \widehat{\mu})' \\ &= \widehat{\Sigma}_1^{-1} \circ \widehat{\Sigma}_1 = \operatorname{trace}(\widehat{\Sigma}_1\widehat{\Sigma}_1^{-1}) \\ &= \operatorname{trace}(I) = d, \end{aligned} \tag{8}$$

where "$\circ$" denotes the trace (element-wise) inner product (i.e. $A \circ B = \text{trace}(A'B)$). Thus, by (8) and the analogous expression for $\widehat{\Sigma}_2$, the likelihood function at $(\widehat{\mu}, \widehat{\Sigma}_1, \widehat{\Sigma}_2)$ can be rewritten as

$$L(\widehat{\mu}, \widehat{\Sigma}_1, \widehat{\Sigma}_2) = (2\pi e)^{-(N_1+N_2)d/2} \det(\widehat{\Sigma}_1)^{-N_1/2} \det(\widehat{\Sigma}_2)^{-N_2/2}. \tag{9}$$

Combining (7) and (9), the original problem of maximizing the likelihood function in μ, Σ_1 and Σ_2 can be reduced to the minimization in $\widehat{\mu}$ of

$$\left(1 + (\bar{X} - \widehat{\mu})' S_1^{-1} (\bar{X} - \widehat{\mu})\right)^{N_1/2} \left(1 + (\bar{Y} - \widehat{\mu})' S_2^{-1} (\bar{Y} - \widehat{\mu})\right)^{N_2/2}, \tag{10}$$

which is the expression Buot and Richards [5] arrived at.

3. Lifting and the EMEP. Expression (10) is already much more tractable than the original likelihood since it depends only on μ. However, the likelihood maximization problem can become substantially more amenable to analysis if it is reformulated as a suitable mathematical programming problem. We can do that by *lifting* it to a higher-dimensional setting, i.e., by including additional variables and constraints, and recasting it in the following way.

Definition 3.1 *The Lifted Gaussian Maximum Likelihood Behrens-Fisher Problem is to solve*

$$\begin{array}{ll} \min\limits_{\mu, u_1, u_2} & f(u_1, u_2) = \frac{N_1}{2} \log(u_1) + \frac{N_2}{2} \log(u_2) \\ & u_1 \geq 1 + (\bar{X} - \mu) S_1^{-1} (\bar{X} - \mu) \\ & u_2 \geq 1 + (\bar{Y} - \mu) S_2^{-1} (\bar{Y} - \mu) \end{array} \tag{11}$$

Since the solutions for the Lifted Gaussian Maximum Likelihood Behrens-Fisher Problem and the original Gaussian Maximum Likelihood Behrens-Fisher Problem must coincide, we will use the acronym BFP to refer to the former from now on.

The advantage to the lifting procedure is to confine the non-convexity of the problem to just two variables, u_1 and u_2. Nevertheless, the objective function still poses a computational challenge since it is non-convex. This means that we can still expect the existence of local solutions as suggested in [5], and further analysis is called for.

One may note, though, that the objective function in (11) is increasing in u_1 and u_2. Moreover, if one of the variables, say u_1, is fixed, then the problem becomes fairly simple: for each value of u_1, we can obtain a solution $u_2^*(u_1)$. The same can be done with u_1^* as a function of u_2. Therefore, associated with (11), we could think of a family of tractable "subproblems" (parameterized

by u_1 for example). Next we will show how to relate the solutions of this family to the solution of the original problem.

Let us focus on the constraints in (11). For a given $\widehat{\mu}$ (a "solution"), consider the squared Mahalanobis distance functions

$$\mathcal{M}_{\bar{X}}(\widehat{\mu}) = (\bar{X} - \widehat{\mu})' S_1^{-1} (\bar{X} - \widehat{\mu}) \quad \text{and} \quad \mathcal{M}_{\bar{Y}}(\widehat{\mu}) = (\bar{Y} - \widehat{\mu})' S_2^{-1} (\bar{Y} - \widehat{\mu}). \tag{12}$$

Note the resemblance between such functions and the generalized distance function G as defined in Kim [15]. They all give ellipsoids in $\widehat{\mu}$, but our use of the functions is different.

Definition 3.2 *The Ellipsoidal Mean Estimation Problem with respect to X at level v_1 is to solve*

$$h_X(v_1) := \min_{\mu} \{\mathcal{M}_{\bar{Y}}(\mu) : \mathcal{M}_{\bar{X}}(\mu) \leq v_1\}, \tag{13}$$

(analogously for Y).

In words, the EMEP with respect to X at level v_1 is to find the *estimate* $\widehat{\mu}_{\text{EMEP}}$ of μ that minimizes the squared distance $\mathcal{M}_{\bar{Y}}$ under the constraint that the squared distance $\mathcal{M}_{\bar{X}}$ is bounded by v_1. The use of the word "estimate" can be justified in at least two ways. First, Gaussian maximum likelihood estimation is based upon finding a vector estimate $\widehat{\mu}_{\text{EMEP}}$ that minimizes a similar quadratic form. Second, the procedure above enjoys the reasonable property that if $\bar{X}$ and $\bar{Y}$ are close (in particular, equal), the solution $\widehat{\mu}_{\text{EMEP}}$ will also be close to $\bar{Y}$ (in particular, equal).

Even though the EMEP is simpler than the BFP, there is no closed-form solution for the former (for given v_1). Nonetheless, EMEP is, in fact, a *convex* problem and can be solved efficiently by a variety of available methods like gradient descent, interior-point methods, cutting-planes, etc. Although all these methods are convergent and a few have good complexity properties (see [3, 12, 18]), in the Appendix we propose a specific algorithm which explores the particular structure of the problem. Not surprisingly, it enjoys better complexity guarantees and better practical performance than the aforementioned methods.

The BFP and the EMEP are in fact closely related. The BFP consists of achieving the optimal balance between the EMEP for X and Y *simultaneously*. This happens because the BFP is based upon the minimization of a function that is monotone in both distance functions. A precise characterization of the relation between the BFP and the EMEP is given in the following theorem.

Theorem 3.1 *Let $(\widehat{\mu}, \widehat{u}_1, \widehat{u}_2)$ be a solution to the BFP. Then, $\widehat{\mu}$ is a solution to the EMEP with respect to X [with respect to Y] at $\widehat{v}_1 = \mathcal{M}_{\bar{X}}(\widehat{\mu})$ [at $\widehat{v}_2 = \mathcal{M}_{\bar{Y}}(\widehat{\mu})$].*

Proof. See Appendix B. ■

Remark 3.1 *Assuming that S_1 and S_2 are positive definite matrices (not only semi-definite), for each level of v_1 the EMEP has a unique solution. However, this does not guarantee that the BFP also has a unique solution, since it could achieve the optimum at two different levels of the distance function.*

4. The Underlying Geometry of the Lifted Behrens-Fisher Problem. In this section we study the nature of the non-convexities in (11), and we show how the feasible set is related to the EMEP. In particular, we obtain a convenient representation of the border of the feasible set that will be used in the algorithm developed in Section 5.

We start by considering the projection of the set of feasible point in (11) into the two-dimensional space of $u = (u_1, u_2)$:

$$\mathcal{K} = \left\{ (u_1, u_2) \in \mathbb{R}^2 : \exists \mu \text{ such that } \begin{array}{l} u_1 \geq 1 + \mathcal{M}_{\bar{X}}(\mu) \\ u_2 \geq 1 + \mathcal{M}_{\bar{Y}}(\mu) \end{array} \right\}. \tag{14}$$

Figure 1 illustrates the geometry of $\mathcal{K}$. Since $\mathcal{M}$ is a convex function, $\mathcal{K}$ is a convex set. Also, $\mathcal{K}$ is unbounded, since $(u_1, u_2) \in \mathcal{K}$ implies that $(u_1 + \gamma_1, u_2 + \gamma_2) \in \mathcal{K}$ as well for arbitrarily values of $\gamma_1, \gamma_2 > 0$. Clearly, $u \in \mathcal{K}$ implies that $u_1 \geq 1$ and $u_2 \geq 1$.

Since the objective function of (11), $f(u) = f(u_1, u_2) = \frac{N_1}{2} \log(u_1) + \frac{N_2}{2} \log(u_2)$, depends only on the variables u, the optimal value of (11) equals

$$\min\{f(u) : u \in \mathcal{K}\}, \tag{15}$$

which still is a non-convex minimization and potentially has many local minima.

However, the representation (15) has two desirable features. First, it completely separates the (non-convex) minimization problem in two variables from the high dimensionality of μ. This will be key to avoid the curse of dimensionality. Second, we can write out a compact region that contains the solution for (15). Define the following problem dependent constants:

$$\begin{array}{rcll} \bar{L}_1 & = & \min_{\mu}\{1 + \mathcal{M}_{\bar{X}}(\mu)\} & = 1 \\ \bar{U}_2 & = & \min_{u_2}\{u_2 : (\bar{L}_1, u_2) \in \mathcal{K}\} & = 1 + \mathcal{M}_{\bar{Y}}(\bar{X}) \\ \bar{L}_2 & = & \min_{\mu}\{1 + \mathcal{M}_{\bar{Y}}(\mu)\} & = 1 \\ \bar{U}_1 & = & \min_{u_1}\{u_1 : (u_1, \bar{L}_2) \in \mathcal{K}\} & = 1 + \mathcal{M}_{\bar{X}}(\bar{Y}) \end{array} \tag{16}$$

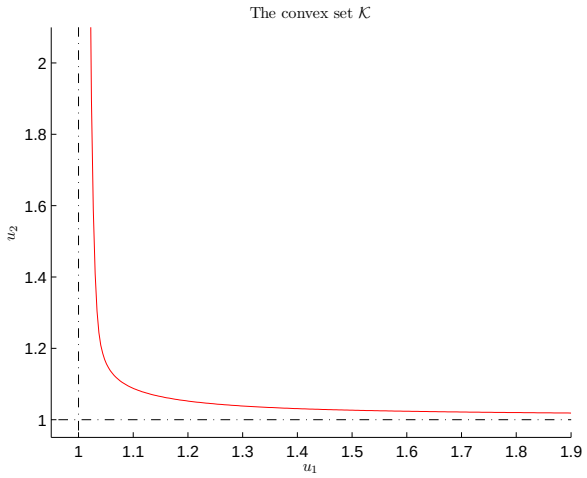

FIG 1. *The convex set $\mathcal{K}$ consists of every point on and above the red curve.*

These quantities define a right triangle

$$\{(\bar{L}_1, \bar{L}_2), (\bar{L}_1, \bar{U}_2), (\bar{U}_1, \bar{L}_2)\} \tag{17}$$

which contains the optimal solution $u^* = (u_1^*, u_2^*)$ for (15). In fact, observe that, by monotonicity, all points in $\mathcal{K}$ above or to the right of the hypotenuse of the triangle have a larger objective value than any point on the hypotenuse. Moreover, the remaining points of $\mathcal{K}$ are contained in the triangle. Therefore the coordinates of the triangle vertices in (17) are lower and upper bounds on the optimal solution (u_1^*, u_2^*), i.e.,

$$\bar{L}_1 \leq u_1^* \leq \bar{U}_1, \quad \bar{L}_2 \leq u_2^* \leq \bar{U}_2.$$

In particular, if $\bar{X} = \bar{Y}$ the triangle degenerates into a single point (as pointed out in [5], the solution is trivial in this case).

Nevertheless, there is a representation cost associated with (15), in the sense that there is no closed-form representation for $\mathcal{K}$ involving only the variables u.

For this reason, we will make use of an additional function g that gives information about (part of) the border of $\mathcal{K}$ (which is where the global optimum is expected to be found, given the quasi-concavity of f). The function g is defined as

$$g(u_1) := \min\{u_2 : (u_1, u_2) \in \mathcal{K}\}.$$

By construction, a point (u_1, u_2) is in $\mathcal{K}$ if and only if $u_2 \geq g(u_1)$. It is easy to show that the function g is convex (its epigraph is exactly the convex set $\mathcal{K}$) and decreasing in u_1.

Note that the function g is directly related to the EMEP with respect to X and the function h_X, since

$$g(u_1) = 1 + \min \quad \begin{array}{ll} \mathcal{M}_{\bar{Y}}(\mu) & = 1 + h_X(u_1 - 1) \\ u_1 - 1 \geq \mathcal{M}_{\bar{X}}(\mu). & \end{array} \tag{18}$$

In other words, evaluating g at u_1 involves solving an EMEP with respect to X.

5. An Algorithm for the Behrens-Fisher Problem. In this section, we propose an algorithm, called Cutting Lines Algorithm (CLA), that generate an ε-solution for the BFP. This means that the algorithm reports a feasible solution at which the objective function value lie within at most ε from the value of the objective function at the optimal solution. Since the feasible solution is given for arbitrary $\varepsilon > 0$, convergence to an optimal solution holds.

The CLA builds upon a polyhedral approximation to the set $\mathcal{K}$. The method optimizes the objective function f over $\hat{\mathcal{K}}_k$ at each iteration. The minimizer point $(u_1, u_2) \in \hat{\mathcal{K}}_k$ is used to improve the polyhedral approximation for the next iteration.

As mentioned in the introduction, it is possible to propose an algorithm based upon the discretization of the range of values of u_1 where we need to evaluate $g(u_1)$. Such algorithm, which we call Discretization Algorithm (DA), can be proved to have better worst-case complexity guarantees than the ones obtained for the CLA. However, Section 6 shows that the practical performance of the CLA strongly dominates that of the DA, since the latter requires evaluating the function g - i.e., solving an EMEP; see expression (18) - at every point of the discretization. Thus, we focus on the CLA and defer the details of the DA to Appendix D.

5.1. *The Cutting Lines Algorithm.* A good way to develop an algorithm for the BFP is to think of constructing sets that (i) approximate $\mathcal{K}$ and (ii) have a simple description involving u. Given the convexity of $\mathcal{K}$, polyhedral approximations to the set $\mathcal{K}$ are a natural candidate. Moreover, such approximations are rather convenient because it is simple to minimize the objective function f over polyhedral sets in two dimensions (see Lemma 5.1 below).

5.1.1. *Building Polyhedral Approximations to $\mathcal{K}$.* Our sequence of polyhedral approximations will be based upon the function g. Given the results for the EMEP, relation (18) implies that for any fixed value of u_1, not only can $g(u_1)$ be efficiently evaluated but also a subgradient $s \in \partial g(u_1)$ (see Lemma C.1 for details) can be easily obtained. Suppose we choose a set of points $\{u_1^i\}_{i=1}^k$ and gather the triples

$$\{u_1^i, g(u_1^i), s^i\}, \quad s^i \in \partial g(u_1^i), \quad i = 1, \ldots, k.$$

By the definition of subgradient, we have that

$$g(u_1) \geq g(u_1^i) + s^i(u_1 - u_1^i) \text{ for all } i = 1, \ldots, k \text{ and } u_1 \in \mathbb{R}.$$

Therefore we can build a minorant polyhedral approximation $\widehat{g}_k$ for g as follows

$$\widehat{g}_k(u_1) = \max_{1 \leq i \leq k} \left\{ g(u_1^i) + s^i(u_1 - u_1^i) \right\}. \tag{19}$$

In turn, such function can be used to build a polyhedral approximation for $\mathcal{K}$ defined as

$$\widehat{\mathcal{K}}_k = \{(u_1, u_2) \in \mathbb{R}^2 : u_2 \geq \widehat{g}_k(u_1)\}.$$

Figure 2 illustrates these relations[1].

The advantage of working with the polyhedral approximation $\widehat{\mathcal{K}}_k$ instead of $\mathcal{K}$ is two-fold. First, $\widehat{\mathcal{K}}_k$ has a much nicer representation (via linear inequalities or extreme points) than $\mathcal{K}$ itself. This is particulary interesting for developing algorithms, which is our goal here. Second, as we anticipated, the minimization of the desired objective function $f(u_1, u_2) = \frac{N_1}{2}\log(u_1) + \frac{N_2}{2}\log(u_2)$ on $\widehat{\mathcal{K}}_k$ is rather tractable, as we show in the following lemma.

Lemma 5.1 *Let $\widehat{\mathcal{K}}_k \subset \mathbb{R}^2_{++}$ be a (convex) polyhedral set. Then the function*

$$f(u_1, u_2) = \frac{N_1}{2}\log(u_1) + \frac{N_2}{2}\log(u_2)$$

is minimized at an extreme point of $\widehat{\mathcal{K}}_k$.

Proof. First, note that since $\widehat{\mathcal{K}}_k \subset \mathbb{R}^2_+$, and because the non-negative orthant is a pointed cone, $\widehat{\mathcal{K}}_k$ must have at least one extreme point. Second, the optimal solution cannot be an interior point of $\widehat{\mathcal{K}}_k$ (otherwise we can strictly decrease both components simultaneously). Third, we recall that f

[1]Such approximation for convex sets can be traced back to the Cutting Planes Algorithm in the Optimization literature [3, 12, 13].

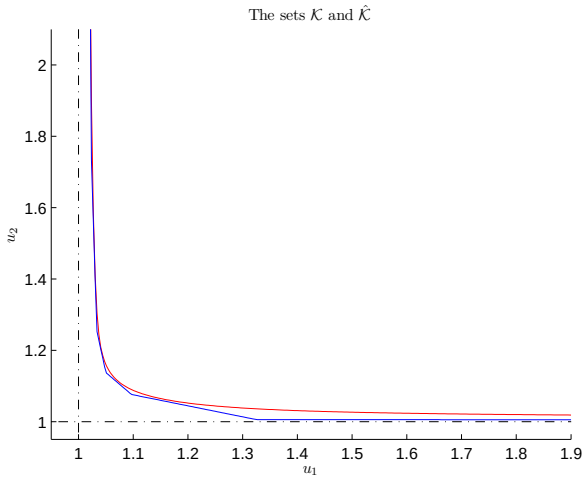

FIG 2. *The convex set $\mathcal{K}$ and its outer polyhedral approximation $\widehat{\mathcal{K}}_k$. The extreme points of $\widehat{\mathcal{K}}_k$ are the kinks of the graph of the piecewise linear function $\hat{g}_k$.*

is a differentiable quasi-concave function. Therefore its gradient is a supporting hyperplane for its upper level sets, which are convex.

Next, suppose that the minimum is achieved at a non-extreme point of $\widehat{\mathcal{K}}_k$, say $x^* = \alpha z + (1-\alpha)y$, for $\alpha \in (0,1)$ and extreme points z, y. By the first order conditions, the gradient of f induces a supporting line for $\mathcal{K}$ at x^* on which both z and y lie. By the (strict) convexity of the upper level sets of f, $\max\{f(z), f(y)\} < f(x^*)$, a contradiction. ■

Since $\widehat{\mathcal{K}}_k$ is an outer approximation of $\mathcal{K}$, minimizing f over $\widehat{\mathcal{K}}_k$ yields a lower bound on the optimal value of (11) for every k. Figure 3 illustrates the minorant approximation of $f(u_1, g(u_1))$ induced by $f(u_1, \hat{g}_k(u_1))$.

5.1.2. *The Algorithm.* The CLA draws upon the minimization of the objective function over the polyhedral approximation $\widehat{\mathcal{K}}_k$ to $\mathcal{K}_k$, which as shown in Lemma 5.1, needs to be carried out only over the extreme points of $\widehat{\mathcal{K}}_k$. A brief description of the algorithm follows. At iteration k, one has a set f^i, $i = 1, ..., k$, of values of the objective function at points (u_1^i, u_2^i), $i = 1, ..., k$, respectively. The values f^i are then compared to $\widehat{f}^k := f(\widehat{u}_1^k, \widehat{u}_2^k)$, where $(\widehat{u}_1^k, \widehat{u}_2^k)$ is the solution of the minimization of f over $\widehat{\mathcal{K}}_k$. If the distance $\min_{0 \le i \le k}(f^i - \widehat{f}^k)$ is small enough (note that $f^i \ge \widehat{f}^k$), the algorithm stops. Otherwise, it takes a new point u_1^{k+1}, slightly to the right of $\widehat{u}_1^k$, and generates

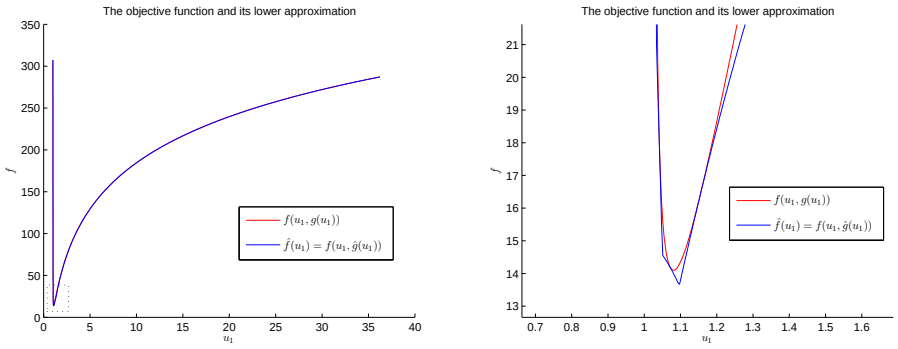

FIG 3. *The outer polyhedral approximation for $\mathcal{K}$ leads to a minorant approximation for f. Therefore lower bounds on the optimal value of (11) are derived if we minimize the minorant approximation $\widehat{f}$. The right figure is a zoom in on the dashed square area of the left figure.*

its corresponding $u_2^{k+1} := g(u_1^{k+1})$ by solving an EMEP. The evaluation of the objective function f at the pair (u_1^{k+1}, u_2^{k+1}) gives a new f^{k+1}, and the algorithm starts over.

Cutting Lines Algorithm (CLA)
Input: Tolerance $\varepsilon > 0$, $u_1^1 = \min\{\bar{U}_1,\ (1+\varepsilon/N_1)\bar{L}_1\}$, $\widehat{g}_0 = 1$, $k = 1$.
Step 1. Evaluate $u_2^k = g(u_1^k)$ and $s^k \in \partial g(u_1^k)$. Compute $f^k = \frac{N_1}{2}\log(u_1^k) + \frac{N_2}{2}\log(u_2^k)$;
Step 2. Define $\widehat{g}_k(u_1) = \max_{0\le i\le k}\{u_2^i + s^i(u_1 - u_1^i)\}$.
Step 3. Compute $\widehat{f}_k = \min\{f(u_1,u_2) : u_2 \ge \widehat{g}_k(u_1), u_1 \ge \bar{L}_1\}$ and the corresponding point $\widehat{u}^k = (\widehat{u}_1^k, \widehat{u}_2^k)$.
Step 4. If $\min_{0\le i\le k}(f^i - \widehat{f}^k) \le \varepsilon$, report $\min_{0\le i\le k} f^i$ and correspondent pair (u_1^{i*}, u_2^{i*}).
Step 5. Else set $u_1^{k+1} \leftarrow \min\{\bar{U}_1,\ \widehat{u}_1^k(1+\varepsilon/N_1)\}$, $k \leftarrow k+1$, and goto Step 1.

Note that each time a new iteration (say, $k+1$) starts, an updated polyhedral approximation $\widehat{\mathcal{K}}_{k+1}$ is constructed through the introduction of a new cut, based on the subgradient $\partial g(u_1^{k+1})$. A new cut removes one extreme

point and creates at most two new extreme points. Therefore, the computational effort of minimizing f over $\widehat{\mathcal{K}}_k$ grows only *linearly* with k (in fact, by keeping track of previous evaluations, re-optimization can be done even faster).

The next theorem shows that the CLA needs only a finite number of iterations to compute a ε-solution.

Theorem 5.1 *The CLA reports an ε-solution for the original problem in at most $\left\lceil \frac{(\bar{U}_1\bar{U}_2)(N_1N_2)}{2\varepsilon^2} \right\rceil$ loops.*

Proof. For $k \geq 1$, note that $u_1^{k+1} \leq \widehat{u}_1^k(1+\varepsilon/N_1)$, and suppose first that $u_2^{k+1} \leq \widehat{u}_2^k(1+\varepsilon/N_2)$. In this case, we have

$$\begin{aligned} f(u_1^{k+1}, u_2^{k+1}) &= \tfrac{N_1}{2}\log(u_1^{k+1}) + \tfrac{N_2}{2}\log(u_2^{k+1}) \\ &\leq \varepsilon + \tfrac{N_1}{2}\log(\widehat{u}_1^k) + \tfrac{N_2}{2}\log(\widehat{u}_2^k) \\ &= \varepsilon + \widehat{f}^k \leq \varepsilon + f^*, \end{aligned}$$

and we have a ε-solution, since (u_1^{k+1}, u_2^{k+1}) is feasible.

Alternatively, if $u_2^{k+1} > \widehat{u}_2^k(1+\varepsilon/N_2)$, we have $u_2^{k+1} > 1$ which implies that $u_1^{k+1} < \bar{U}_1$. Therefore $u_1^{k+1} = \widehat{u}_1^k(1+\varepsilon/N_1)$ and the next Cutting Lines approximation removes at least a rectangle of area $\frac{\varepsilon^2}{N_1N_2}\widehat{u}_2^k\widehat{u}_1^k$ between the difference of $\widehat{\mathcal{K}}_k$ and $\mathcal{K}$. Since the area difference between these sets was bounded by $\bar{U}_1\bar{U}_2/2$ at the very first iteration, the algorithm performs at most

$$\left\lceil \frac{(\bar{U}_1\bar{U}_2)(N_1N_2)}{2\varepsilon^2} \right\rceil$$

loops. ■

This computational complexity result immediately yields the following convergence results.

Corollary 5.1 *For $\varepsilon_k \downarrow 0$, let (u_1^k, u_2^k) be the ε_k-solutions to (15) and let the vectors (μ^k, u_1^k, u_2^k) be their induced ε_k-solutions to the (Lifted) BFP. Then, every accumulation point of the sequence $\{(\mu^k, u_1^k, u_2^k)\}_{k\in\mathbb{N}}$ is a solution to the BFP.*

Corollary 5.2 *The CLA can be used to generate a sequence of points that converge to a global solution of the (Lifted) Behrens-Fisher Problem.*

5.2. *Computational Experiments with CLA and DA.* Our complexity bound for the CLA is worse than that for the DA. However, the DA solves the EMEP for every point of the discretized domain of u_1. In contrast, the

CLA seeks to produce a certificate of ε-optimality at each iteration by comparing the best current solution and the solution to the minimization on $\widehat{\mathcal{K}}_k$. In computational practice, this drastically reduces the number of necessary iterations to find an ε-solution, as can be seen in Table 1.

Table 1 reflects the expected computational behavior of the methods. As the dimension increases, more effort is needed but the CLA is order of magnitudes faster than the DA, since the latter requires the complete discretization of the interval $[\overline{L}_1, \overline{U}_1]$. Such requirement of evaluating the function g on $O(1/\varepsilon)$ different points (remember that the complexity analysis is exact in the case of the DA) seems to be a naive approach, indeed.

Medium Size Instances			Average Running Times (in seconds)			Average Iterations	
d	N_1	N_2	Initialization	DA	CLA	DA	CLA
20	100	200	0.01	6.89	0.01	6853.2	15.4
30	150	300	0.02	18.97	0.02	10859.6	17.5
40	200	400	0.03	12.92	0.04	9256.8	17.4
50	250	500	0.05	14.37	0.03	8414.6	18.5
60	300	600	0.08	25.85	0.04	10495	17.7
70	350	700	0.13	27.93	0.05	8502.1	17.6
80	400	800	0.18	44.21	0.08	9912.7	18.7
90	450	900	0.24	68.38	0.10	11796.4	18.3
100	500	1000	0.32	74.98	0.13	9859.5	19.0

Large Size Instances			Average Running Times (in seconds)			Average Iterations	
d	N_1	N_2	Initialization	DA	CLA	DA	CLA
200	1000	2000	2.07	-	0.81	-	20.8
300	1500	3000	6.64	-	2.68	-	19.8
400	2000	4000	16.08	-	6.31	-	20.1
500	2500	5000	43.35	-	13.16	-	21.2
600	3000	6000	56.62	-	24.46	-	21.5
700	3500	7000	87.88	-	39.54	-	22.0
800	4000	8000	142.05	-	49.73	-	20.9
900	4500	9000	455.23	-	91.05	-	22.1
1000	5000	10000	671.80	-	154.57	-	22.3

TABLE 1

Computational times (in seconds) and total number of iterations (which equal the number of EMEP problems solved) of the computational experiments with relative tolerance $\varepsilon = 10^{-3}$.

The polyhedral approximation used in the CLA provides a way of focusing the search on a promising region, a concept well exploited in the Optimization literature. Table 1 also illustrates the number of loops required by each algorithm in the test problems.

The number of loops performed by the Discretization Algorithm depends

only on the precision ε, and on the problem dependent values of $\bar{L}_1$ and $\bar{U}_1$. On the other hand, these problem dependent quantities do not seem to affect the CLA. This points to the question of whether there exists a (better) complexity analysis for the CLA which might be independent of these quantities.

The implementation of the algorithms is an easy task in any programming package where matrix inversion and spectral decomposition subroutines for positive definite matrices are available (e.g., R, Matlab, etc.). We do not claim to have the most efficient implementation of the methods proposed here. Nevertheless, our numerical results show that the CLA is computationally efficient and scales quite nicely as the data dimension d increases. The underlying reason is the certificate of optimality that the method is constructing on each iteration. The value $\widehat{f}_{min}$ provides a lower bound for the optimal solution which is used to construct a stopping criterion. For a problem whose dimension is greater than one thousand, numerical approximations on the computation of the spectral decomposition are a potential limitation for the method proposed in the Appendix C. An alternative approach is to compute an inverse matrix at each iteration of the EMEP, which will lead to a more robust implementation at the cost of additional running time (see [6] for details).

In our experiments we use medium and large size instances where the data dimension d varies from 20 to 1000. The results were generated using a relative precision of $\varepsilon = 10^{-3}$. We report the average over ten different instances. The DA has proved to be too cumbersome for large instances.

6. Finite Sample Properties of the Wald, Likelihood Ratio, and Lagrange Multiplier Tests through the CLA. Three commonly used multivariate tests based upon the maximization of the likelihood function $L(\mu, \cdot, \cdot)$ as defined in (9) are the Wald (W), Likelihood Ratio (LR), and the Lagrange Multiplier (LM) Tests. For a certain hypothesized restriction on the parameter space

$$H_0 : c(\mu) = q,$$

the test statistics are defined as

$$W = [c(\widehat{\mu}) - q]'(\mathrm{Var}(c(\widehat{\mu}) - q))^{-1}[c(\widehat{\mu}) - q]$$

$$LR = \mathrm{argmax}\frac{L_R(\mu)}{L(\mu)},$$

where $L_R(\mu) =$ likelihood function under H_0, and

$$LM = \mathrm{argmax}[L(\mu) + \lambda'(c(\mu) - q)],$$

where λ is a vector of Lagrange multipliers.

The W, LR and LM Tests are asymptotically equivalent under the null hypothesis. However, their behavior can be rather different in small samples, and their finite sample properties are usually unknown, except for a few particular cases (see, for instance, Greene [11] and Godfrey [10]). In this section, we use the CLA to investigate and compare the finite sample properties - size and power - of these tests. In particular, we are interested in the sensitivity of the tests to dimensionality.

We emphasize that the CLA allows for the study of the properties of the tests in high-dimensional contexts. In contrast, the literature on the BFP typically overlooks the issue and reports results for small dimensional problems, typically smaller than $d = 6$ and in general no greater than $d = 10$.

6.1. *Conflict Among Criteria.* It is well-known that the W, LR, and LM statistics for testing linear restrictions in the context of classical linear models satisfy the inequalities $W \geq LR \geq LM$ (see Savin [21], Berndt and Savin [2], Breusch [4] and Godfrey [10]). Before turning to simulations, we show that such inequalities also hold in the case of the BFP.

Theorem 6.1 *For the BFP,*

$$W \geq LR \geq LM.$$

Proof. To show the first inequality note that, using since $\log(1+\delta) \leq \delta$, we have

$$LR \leq c^* = \min_{\mu} N_1(\bar{X}-\mu)S_1^{-1}(\bar{X}-\mu) + N_2(\bar{Y}-\mu)S_2^{-1}(\bar{Y}-\mu).$$

The optimal solution of the right hand side is achieved at $\mu_{c^*} = (N_1S_1^{-1} + N_2S_2^{-1})^{-1}(N_1S_1^{-1}\bar{X} + N_2S_2^{-1}\bar{Y})$. Using μ_{c^*}, and the matrix identities

$$(A+B)^{-1} = A^{-1} - A^{-1}(A^{-1}+B^{-1})^{-1}A^{-1} = A^{-1}(A^{-1}+B^{-1})^{-1}B^{-1},$$

we prove that $c^* = (\bar{X}-\bar{Y})'(S_1/N_1 + S_2/N_2)^{-1}(\bar{X}-\bar{Y}) = W$.

Let $\hat{\mu}$ be a solution for the BFP. After simplifications, the LM statistic can be written as

$$LM = N_1(\bar{X}-\hat{\mu})'\widehat{\Sigma}_1^{-1}(\bar{X}-\hat{\mu}) + N_2(\bar{Y}-\hat{\mu})'\widehat{\Sigma}_2^{-1}(\bar{Y}-\hat{\mu}).$$

Next note that

$$(\bar{X}-\hat{\mu})'\widehat{\Sigma}_1^{-1}(\bar{X}-\hat{\mu}) = (\bar{X}-\hat{\mu})'S_1^{-1}(\bar{X}-\hat{\mu}) - \frac{\left[(\bar{X}-\hat{\mu})'S_1^{-1}(\bar{X}-\hat{\mu})\right]^2}{1+(\bar{X}-\hat{\mu})'S_1^{-1}(\bar{X}-\hat{\mu})}$$

by using a rank-one update formula[2] for $\widehat{\Sigma}_1^{-1}$. The result follows by considering the term for Y as well and noting that $\log(1+\delta) \geq \delta - \frac{\delta^2}{1+\delta}$. ■

6.2. *Monte Carlo Study of the Size of the Test.* Inequalities (6.1) imply that the rejection rate of the W Test is greater than or equal to that of the LR Test, which in turn is greater than or equal to that of the LM Test. A more accurate understanding of the extent to which this influences the size and the power of such tests can be obtained through simulations.

We performed a Monte Carlo study of the finite-sample properties of the W, LR, and LM tests at sizes $\alpha = 0.01, 0.05, 0.10$. The rejection regions were defined based upon Wilks' Theorem on the asymptotic χ^2_d distribution of the test statistic.

The study also includes the Likelihood Ratio statistic with the Bartlett correction

$$B := \Big(1 - \frac{\widehat{c}_1}{N-2}\Big) LR,$$

where

$$\widehat{c}_1 = \frac{\widehat{\psi}_1 - \widehat{\psi}_2}{d},$$

$$\widehat{\psi}_1 = \frac{N_2^2(N-2)}{N^2(N_1-1)}\{\mathrm{tr}(S_1\overline{S}^{-1})\}^2 + \frac{N_1^2(N-2)}{N^2(N_2-1)}\{\mathrm{tr}(S_2\overline{S}^{-1})\}^2,$$

$$\widehat{\psi}_2 = \frac{N_2^2(N-2)}{N^2(N_1-1)}\{\mathrm{tr}(S_1\overline{S}^{-1}S_1\overline{S}^{-1})\} + \frac{N_1^2(N-2)}{N^2(N_2-1)}\{\mathrm{tr}(S_2\overline{S}^{-1}S_2\overline{S}^{-1})\},$$

and $\overline{S} = \frac{N_2}{N}S_1 + \frac{N_1}{N}S_2$.

The Bartlett correction as defined above provides an $O(N^{-2})$ approximation to the mean of the χ^2_d distribution (more details can be found in Yanagihara and Yuan [29]). We will refer to the LR Test under the Bartlett correction as the B Test.

To facilitate comparison with other works on the multivariate BFP (e.g. Yao [30], Subrahmaniam and Subrahmaniam [26], Kim [15], Krishnamoorthy and Yu [16]), we performed tests for the low dimensional cases of $d = 2, 5$ and 10, but we also included the higher-dimensional cases of $d = 25, 50, 75$, 100 and 200. For each d, the sample sizes used were $N_1 = 5d, 10d, 20d$, and $N_2 = 2N_1$. For a given dimension size d, each covariance matrix Σ_i, $i = 1, 2$, was constructed by creating an initial matrix S_i with $N(0,1)$ entries, and then setting $\Sigma_i = S_i S_i'$.

[2]For invertible M and a vector v, the inverse of the rank-one update of M by vv' can be written as $(M + vv')^{-1} = M^{-1} - \frac{M^{-1}vv'M^{-1}}{1+v'M^{-1}v}$.

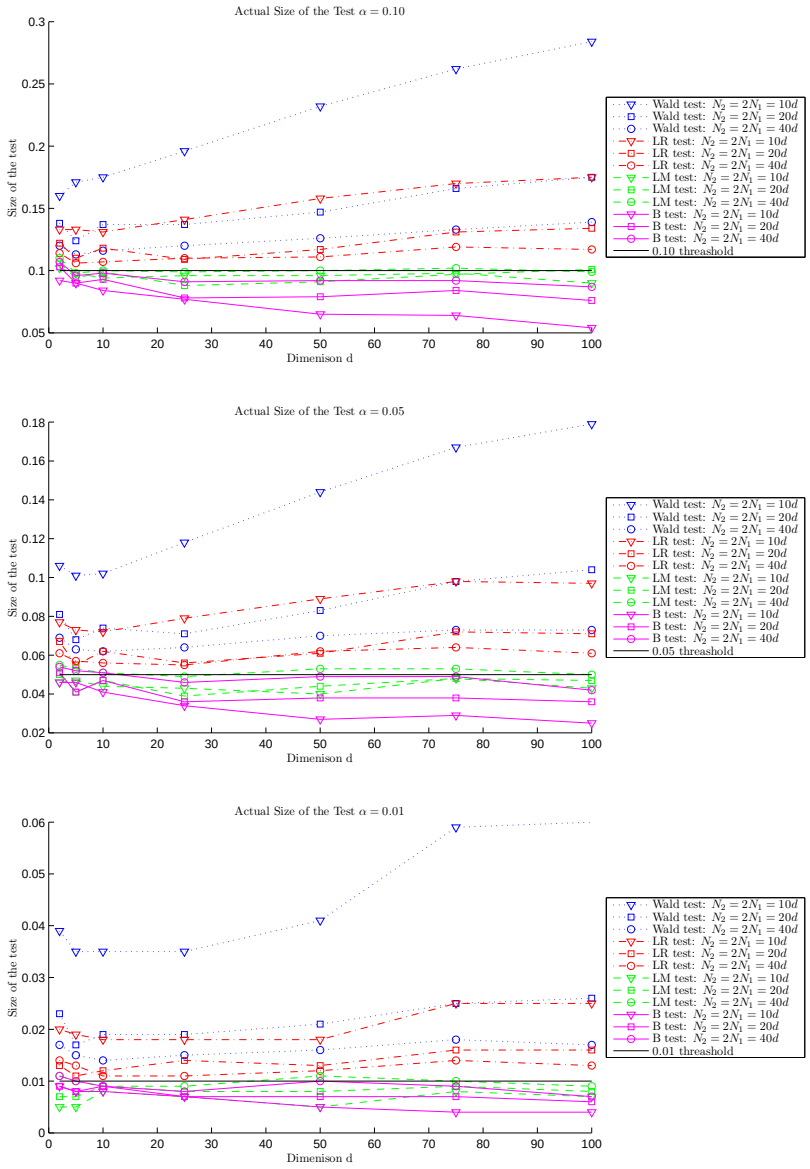

FIG 4. *The behavior of the size of the tests when the dimension increases and the ratio between the number of observations and dimension is fixed.*

The results can be seen in Figure 4 (the actual numerical output can be found in Table E in the appendix). Each entry was generated using 10,000 runs. The W and the LR tests tend to over-reject the null hypothesis while the LM Test tends to slightly under-reject it. We kept constant the ratio between the number of observations and the dimension so that we can observe how the quality of the approximation behaves as the dimensionality of the problem grows. One may notice how sensitive the W and the LR Tests are to increases in the dimension. Only for the (relatively) large sample case $N_1 = 20d$ does the LR Test have actual size fairly close to α. On the other hand, the W Test appears to demand even (relatively) larger samples. For instance, when $d = 100$ and $\alpha = 0.10$, even when $N_1 = 20d$ the W test is off by 3.8 percentage points.

In contrast with the W and the LR Tests, the LM shows remarkable robustness with respect to dimensionality. For all α, there does not appear to be any clear (say, monotonic) pattern of change on the actual test size with respect to increases in dimensionality, or maybe even sample size N_1.

For all values of α and different sample sizes, the B Test is roughly as accurate as the LM Test for low dimensional settings (roughly, $d \leq 20$). For $d > 20$, though, it grossly over-compensates the over-rejection rates of the W Test, with the possible exception of the comparatively large sample sizes $N_1 = 20d$.

Figure 4 illustrates the above comments. Accordingly, the W Test usually shows the steepest curve of dimension vs actual test size for different N_1, while the LM Test displays approximately horizontal curves, especially for higher-dimensional settings.

6.3. *Monte Carlo Study of the Power of the Test.* We performed computational experiments on the power of the W, LR, LM and B Tests for the cases of dimension $d = 10, 50, 100$, and sample sizes $N_1 = 5d$, $10d$, and $20d$, with $N_2 = 2N_1$.

The analysis of the power for multivariate tests is naturally more difficult due to the multi-dimensionality of the parameter space. For this reason, we chose to investigate and compare the power of the W, LR, LM and B Tests over a standardized parameter space in the following sense. For each simulation run, covariance matrices Σ_1 and Σ_2 were (randomly) generated through the same procedure as the one for the evaluation of the sizes of the test. The mean of X, μ_1, was set to zero by default. The choice of the mean(s) of Y, $\mu_2(\Delta)$, was made as solution(s) to the squared Mahalanobis distance equation(s)

$$(\mu_1 - \mu_2(\Delta))'(\Sigma_1 + \Sigma_2)^{-1}(\mu_1 - \mu_2(\Delta)) = \Delta^2,$$

where Δ represents a family of appropriately selected constants. For convenience, such solutions $\mu_2(\Delta)$ were always taken on some canonical axis, and the specific axis chosen changed across simulation runs. The use of randomly standardized Mahalonobis distances is justified by the fact that the BFP is defined without information on the population covariances.

The results are depicted in Figure 4, which contains plots for dimensions $d = 10$, 50 and 100. Colors represent tests, while geometric figures represent sample sizes (e.g., a triangle symbolizes $N_1 = 5d$).

Perhaps the most striking feature of all four plots ($d = 10$, 50 and 100) is the fact that, for a given sample size N_1, the *shapes* of the power curves for the four tests look alike. More specifically, given N_1, the curve for the W Test looks like an up-shifted version of the curve for the LR Test, which in turn looks like an up-shifted version of the curve for the LM Test. The same is true for the curve for the B Test, which lies mostly below the curve for the latter. The observed "order" of the curves should not come as a surprise. First, regarding the W, LR and LM Tests, because of the theoretical inequalities in Theorem 6.1. Second, because the simulation results for the test sizes show that the W and LR Tests tend to over-reject the null hypothesis (the former, substantially more than the latter), while the LM Test has size close to α and the B Test tends to under-reject the null hypothesis. In other words, we are essentially comparing tests of *different sizes* (see also the conclusions in Breusch [4] for the case of linear regression). The shape of the curves suggests the possibility that, if test size adjustment is made for the W and LR Tests, the power curves of the three tests may get rather close to each other. Such adjustment would imply, of course, going beyond Wilks' Theorem and developing exact quantiles, especially for the W and the LR Tests.

The plot for the low-dimensional case of $d = 10$ displays a "well-behaved" pattern, in the sense that the curves for different tests and for the same sample size tend to be grouped together. In particular, the curves for sample size $N_1 = 40d$ are almost super-imposed, which means that, power-wise, the tests are nearly equivalent in this situation. Note that the curves for sample size $N_1 = 10d$ (triangle) lie above the remaining ones close to the origin, i.e., in the case where the Mahalanobis distance between μ_1 and μ_2 is small. Again, this should not come as a surprise, since the simulation results for the test sizes (i.e., zero Mahalanobis distance between μ_1 and μ_2) show that relatively small sample sizes imply a tendency for over-rejection in the case of the W and LR Tests.

The effect of higher dimensionality can be seen in the two remaining plots ($d = 50$ and 100). The main impact seems to be greater vertical distances

among the curves for the four tests, particularly for the cases of smaller sample sizes. Even for the higher-dimensional case $d = 100$, though, the larger sample size $N_1 = 20d$ brings the curves a lot closer to each other. As one might expect, larger sample sizes compensate for high dimension and point to the asymptotic equivalence of the W, LR, LM and B Tests.

7. Extension to Behrens-Fisher-like Problems. It should be noted that the methodology proposed in this paper can be applied to a much broader class of problems. Strictly speaking, all we need is to be able to replicate the strategy of constructing lifted problems whose solution lie on extreme points of a two dimensional convex domain[3], and to evaluate the subproblems which define the convex domain. A sufficient condition for this is the quasi-concavity of the objective function of the lifted problem and the convexity of the subproblems.

To set up a broader framework, assume we have two random samples $\{X_i\}_{i=1}^{N_1}$ and $\{Y_i\}_{i=1}^{N_2}$ whose log-likelihood functions are denoted by $l_1(X;\mu,\alpha)$ and $l_2(Y;\mu,\beta)$, respectively. The generalized M-estimation problem of interest is defined as

$$\max_{\mu,\alpha,\beta} \; l_1(X;\mu,\alpha) + l_2(Y;\mu,\beta).$$

A generalization of the subproblem can be cast in terms of the log-likelihood functions directly. Assume there exist two monotone (decreasing) transformations $T_X, T_Y : \mathbb{R} \to \mathbb{R}$ such that $T_X(l_1(X;\cdot,\cdot))$ and $T_Y(l_2(Y;\cdot,\cdot))$ are convex functions. The subproblems, analogous to the EMEP, are

$$h_X(u_1) = \min_{\mu,\alpha,\beta} \left\{ T_Y\left(l_2(Y;\mu,\beta)\right) : T_X\left(l_1(X;\mu,\alpha)\right) \leq u_1 \right\}.$$

The geometric results in Section 4 still hold with minor modifications. Moreover, under the above convexity assumption, the evaluation of $h_X(u_1)$ can be efficiently performed through standard convex programming techniques. Therefore, the convergence results of Section 5 are still valid.

The above framework encompasses the BFP by taking $T_X(z) = \exp(\frac{2}{N_1}z) - 1$ and $T_Y(z) = \exp(\frac{2}{N_2}z) - 1$. The lifted problem is given by equation (11), whose objective function is concave, and solving the subproblem $h_X(u_1)$ is equivalent to solving an EMEP.

We now give a simple example of the application of the methodology described above to a Behrens-Fisher-like problem.

[3]Higher-dimensional convex domains would impose an additional burden in terms of computational complexity.

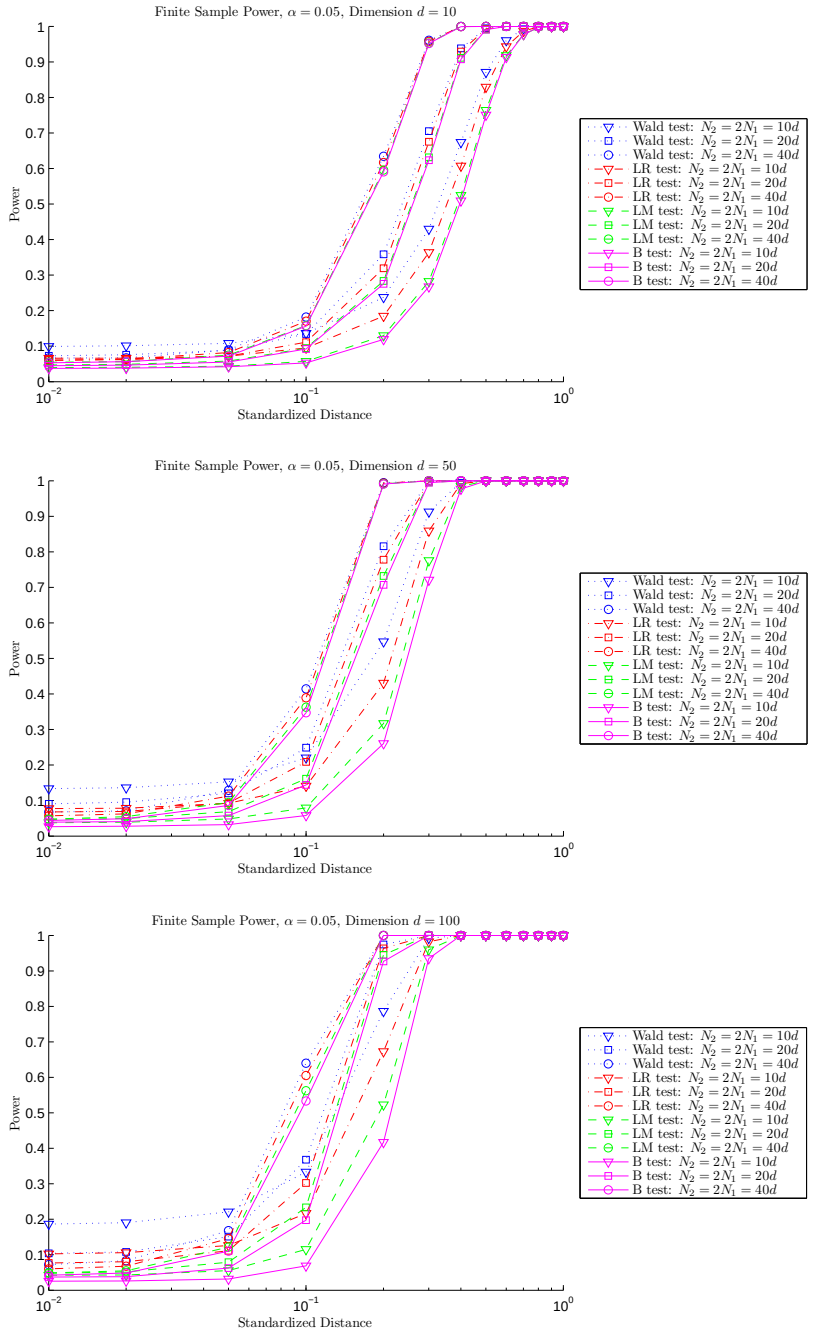

FIG 5. *Monte Carlo study of the power of the W, LR, LM and B Tests for the size* $\alpha = 0.05$ *with different sample sizes and dimensions equal to* 10, 50, *and* 100.

Example 7.1 *Assume $X \sim N(\mu, \Sigma)$ but, differently from the BFP, Y follows a multivariate Laplacian distribution, i.e.,*

$$f_Y(y) = c_L \exp(-||y - \mu||),$$

where c_L is the normalization constant and $\|\cdot\|$ is the Euclidean norm. The related lifted problem can be cast as

$$\begin{array}{ll} \min\limits_{\mu, u_1, u_2} & f(u_1, u_2) = \frac{N_1}{2}\log(u_1) + u_2 \\ & u_1 \geq 1 + \mathcal{M}_{\bar{X}}(\mu) \\ & u_2 \geq \sum_{i=1}^{N_2} \|Y_i - \mu\|. \end{array} \tag{20}$$

Here, the problem objective function is concave (u_1, u_2), and therefore the solution must lie on the border of the convex domain of these variables. Such domain can be written as

$$\mathcal{K} = \left\{ (u_1, u_2) \in \mathbb{R}^2 : \exists \mu \text{ such that } \begin{array}{l} u_1 \geq 1 + \mathcal{M}_{\bar{X}}(\mu) \\ u_2 \geq 1 + \sum_{i=1}^{N_2} \|Y_i - \mu\| \end{array} \right\}. \tag{21}$$

Moreover, the associated subproblems, using $T_X(z) = \exp(\frac{2}{N_1}z) - 1$ and $T_Y(z) = z$, are convex programming problems and have the form

$$h_X(u_1) = \min_{\mu} \left\{ \sum_{i=1}^{N_2} \|Y_i - \mu\| \ : \ \mathcal{M}_{\bar{X}}(\mu) \leq u_1 \right\} \quad \text{and}$$

$$h_Y(u_2) = \min_{\mu} \left\{ \mathcal{M}_{\bar{X}}(\mu) \ : \ \sum_{i=1}^{N_2} \|Y_i - \mu\| \leq u_2 \right\}.$$

Both these problems can be solved via convex quadratic programming, which can be done quite efficiently even in high-dimensional cases.

APPENDIX A: NOTATION OF CONVEX ANALYSIS

Herein we gather the definitions of relevant concepts in Convex Analysis for this work. We refer to [20] for a analytic exposition of Convex Analysis and to [12] for a more geometric one.

A set S is convex if for any $x,\ y \in S$, $\alpha \in [0, 1]$, $\alpha x + (1 - \alpha)y \in S$. An extreme point of a convex set is a point that cannot be written as a strictly ($\alpha < 1$) convex combination of any other distinct points in the set. A set P is said to be polyhedral if $P = \{x \in \mathbb{R}^n : Ax \leq b\}$, where A is a matrix and b a vector. It follows that polyhedral sets are convex and their extreme

points are its corners. The recession cone C_S of a convex set S is the set of directions that go to infinity in S, formally, $C_S = \{d : d + S \subset S\}$.

A function $g : \mathbb{R}^n \to \mathbb{R}$ is said to be convex if for any $x, y \in \mathbb{R}^n$, and $\alpha \in [0,1]$, $g(\alpha x + (1-\alpha)y) \leq \alpha g(x) + (1-\alpha)g(y)$. A function $f : \mathbb{R}^n \to \mathbb{R}$ is quasi-concave if for any $x, y \in \mathbb{R}^n$, and $\alpha \in [0,1]$, $f(\alpha x + (1-\alpha)y) \geq \min\{f(x), f(y)\}$, or equivalently, the upper level sets of f are convex sets.

Associated with a convex function $g : \mathbb{R}^n \to \mathbb{R}$, we can define its subdifferential at x as $\partial g(x) = \{s \in \mathbb{R}^n : g(y) \geq g(x) + \langle s, y - x\rangle, \text{ for all } y \in \mathbb{R}^n\}$. The elements of the subdifferential, also called subgradients, play the role of the gradient in case g is nondifferentiable. Note that $\partial g(x)$ is always non-empty.

APPENDIX B: THE RELATION BETWEEN THE BFP AND THE EMEP

Proof of Theorem 3.1. Without loss of generality, we will develop the argument only for the EMEP with respect to X.

Let $\widehat{\mu}_{\text{EMEP}}$ be a solution to the EMEP with respect to X at some positive v_1. By the monotonicity of log, this means that the expression

$$\frac{N_1}{2}\log(1 + v_1) + \frac{N_2}{2}\log(1 + \mathcal{M}_{\bar{Y}}(\mu)) \tag{22}$$

is minimized at $\widehat{\mu}_{\text{EMEP}}$.

Now, let $(\widehat{\mu}, \widehat{u}_1, \widehat{u}_2)$ be a solution to the BFP problem. This means that the expression

$$\frac{N_1}{2}\log(1 + \mathcal{M}_{\bar{X}}(\mu)) + \frac{N_2}{2}\log(1 + \mathcal{M}_{\bar{Y}}(\mu)) \tag{23}$$

is minimized at $\widehat{\mu}$ and we have $\widehat{u}_1 = 1 + \mathcal{M}_{\bar{X}}(\widehat{\mu})$. Since expression (23) is an upper bound for expression (22) when we set $v_1 := \mathcal{M}_{\bar{X}}(\widehat{\mu})$, $\widehat{\mu}$ is also a solution to the EMEP with respect to X at v_1. ■

APPENDIX C: SOLVING THE EMEP

Consider the convex problem in (13). There are a variety of "general purpose" convergent algorithms that can solve it. Here, we propose a specific algorithm tailored for the particular structure of the EMEP.

Let λ be the (nonnegative) LM associated with the inequality constraint. The first order conditions are necessary and sufficient, and are given by

$$2S_2^{-1}(\bar{Y} - \mu) + 2\lambda S_1^{-1}(\bar{X} - \mu) = 0, \quad \text{and}$$

$$\lambda(\mathcal{M}_{\bar{X}}(\mu) - v_1) = 0.$$

Assuming that $\lambda > 0$ (otherwise, the solution is just $\widehat{\mu} = \bar{Y}$), the optimal $\widehat{\mu}$ is a function only of λ:

$$\widehat{\mu}(\lambda) = (S_2^{-1} + \lambda S_1^{-1})^{-1}(S_2^{-1}\bar{Y} + \lambda S_1^{-1}\bar{X}). \tag{24}$$

Therefore, in order to solve the EMEP, it suffices to compute a root λ^* of the following non-linear univariate function:

$$m(\lambda) = \mathcal{M}_{\bar{X}}(\widehat{\mu}(\lambda)) - v_1. \tag{25}$$

The algorithm we propose here is based upon the algorithm proposed by Ye [31], who in turn built upon earlier work by Smale [24].

Our algorithm has two main parts. The first part consists of a binary search over intervals of increasing length to find which interval I_{i^*} contains what Smale [24] calls an *approximate root*:

Definition C.1 *A point λ^0 is said to be an approximate root of an analytic real function $m : \mathbb{R} \to \mathbb{R}$ if*

$$|\lambda^{k+1} - \lambda^k| \leq (1/2)^{2^{k-1}-1}|\lambda^1 - \lambda^0|.$$

In the second part of the algorithm, Newton's method is used over the interval I_{i^*} to find the approximate root λ^*. For the sake of exposition, we focus on the case of $m : \mathbb{R} \to \mathbb{R}$ (although the results in [31] hold in much greater generality). Recall that the Newton iterate for a function m from a current point λ^k is:

$$\lambda^{k+1} = \lambda^k - \frac{m(\lambda^k)}{m'(\lambda^{k+1})}.$$

Newton's method converge quadratically from the very first iteration. In [24], Smale gives sufficient conditions under which a particular point is an approximate root. Although it is hard to verify Smale's condition in general, Ye provided a constructive method to find such a point for a particular class of functions. Ye's results in [31] apply in our case. Our algorithm is as follows.

Binary Search and Newton Method

Input: Upper and lower bounds on the value of the root $[a, b]$, $b \geq a \geq \delta$, tolerance $\delta > 0$.

Step 1. Define a partition of [a,b] through intervals of the form $I_i = [a(1 + 1/12)^i, a(1 + 1/12)^{i+1})$.

Step 2. Perform binary search on these intervals to find I_{i^*} that contains the true root λ^*.

Step 3. Let $\lambda^0 = a(1 + 1/12)^i$, $k = 0$.

Step 4. Perform Newton's method from λ^k: $\lambda^{k+1} \leftarrow \lambda^k - \frac{m(\lambda^k)}{m'(\lambda^k)}$

Step 5. Stop if $k > 1 + \log_2 (1 + \max\{0, \log_2(b/\delta)\})$ steps.

Step 6. Else set $k \leftarrow k + 1$, and goto Step 4.

Theorem C.1 *After the computation of a spectral decomposition of the matrix $S_1^{1/2} S_2^{-1} S_1^{1/2}$, and given a desired precision $\delta > 0$ and an upper bound b for the solution, the algorithm finds a δ-approximate solution $\widehat{\lambda}$ such that $|\lambda^* - \widehat{\lambda}| < \delta$ in at most*

$$O\left(d \log\log \frac{b}{\delta}\right)$$

arithmetic operations.

Proof. Making the following change of variables/notation

$$w := S_1^{-1/2}(\mu - \overline{X}), \quad M := S_1^{1/2} S_2^{-1} S_1^{1/2} = PDP^T,$$

$$v = 2S_2^{-1} S_1^{1/2}(\bar{Y} - \bar{X}), \quad \text{and} \quad s = P^T v$$

problem (18) is equivalent to

$$\begin{aligned} h(v_1) = \min \quad & w^T M w - v^T w \\ & \|w\|^2 \leq v_1 \end{aligned}$$

up to a constant value (which does not matter for the optimization).

Under the new notation we can rewrite the function m as

$$m(\lambda) = s^T (D + \lambda I)^{-2} s - v_1 = \sum_{i=1}^{d} \frac{s_i^2}{(D_i + \lambda)^2} - v_1.$$

The function $m(\lambda)$ is analytic and its derivatives can be easily computed as

$$m^{(k)}(\lambda) = (-1)^k (k+1)! \sum_{i=1}^{d} \frac{s_i^2}{(D_i + \lambda)^{k+2}}.$$

Note that $m' < 0$ and $m'' > 0$ (i.e., m is decreasing and convex). Thus, we can evaluate m and m' in $O(d)$ operations. This implies that each Newton step can be implemented in $O(d)$ arithmetic operations.

Let $\lambda^0 = a(1 + 1/12)^{i^*}$ be the left endpoint of the interval selected by binary search. From Ye [31] it follows that λ^0 satisfies Smale's sufficient condition to be an approximate root. Therefore Newton's Method converges quadratically from the very first iteration from λ^0. From the convexity of m, the convergence is monotone, i.e., $0 < \lambda^0 < \lambda^k < \lambda^{k+1} < \lambda^* \leq b$ for every k (in particular, we have $|\lambda^1 - \lambda^0| < b$). This implies that we need at most

$$k = 1 + \log_2(1 + \max\{0, \log_2(b/\delta)\})$$

Newton steps to achieve a $|\lambda^k - \lambda^*| < \delta$. Moreover, the total number of subintervals is $\frac{1}{\log(1+1/12)} \log(b/a)$. The binary search can thus be implemented in $O(\log\log(b/a))$. The result follows by noting that we can take $a \geq \delta$. ■

Remark C.1 *Even when we need to solve the EMEP for many different levels of the Mahalanobis distance function, the spectral decomposition of $S_1^{1/2} S_2^{-1} S_1^{1/2}$ needs to be performed only once. This feature of the algorithm makes it a good auxiliary method for the CLA, described in Section 5.*

The following lemma illustrates how to obtain subgradients for the function h_X with no additional computational effort, which is of interest for the CLA.

Lemma C.1 *Let λ^* be a root of the function m as defined in (25). Then $-\lambda^*$ is a subgradient of h_X at v_1.*

Proof. Recall $m(\lambda^*) = 0$ implies that $\mu(\lambda^*)$ minimizes $\mathcal{M}_{\bar{Y}}(\mu) + \lambda^* \mathcal{M}_{\bar{X}}(\mu)$. For any v we have

$$\begin{aligned} h_X(v_1) &= \mathcal{M}_{\bar{Y}}(\widehat{\mu}(\lambda^*)) = \mathcal{M}_{\bar{Y}}(\widehat{\mu}(\lambda^*)) + \lambda(\mathcal{M}_{\bar{X}}(\widehat{\mu}(\lambda^*)) - v_1) \\ &= \mathcal{M}_{\bar{Y}}(\widehat{\mu}(\lambda^*)) + \lambda^*(\mathcal{M}_{\bar{X}}(\widehat{\mu}(\lambda^*)) - v) + \lambda^*(v - v_1) \\ &\leq h_X(v) + \lambda^*(v - v_1) \end{aligned}$$

where we used weak duality ($\min\max \geq \max\min$) as follows

$$\begin{aligned} h_X(v) &= \min_\mu \max_{\lambda \geq 0} \mathcal{M}_{\bar{Y}}(\mu) + \lambda(\mathcal{M}_{\bar{X}}(\mu) - v) \\ &\geq \max_{\lambda \geq 0} \min_\mu \mathcal{M}_{\bar{Y}}(\mu) + \lambda(\mathcal{M}_{\bar{X}}(\mu) - v) \\ &\geq \mathcal{M}_{\bar{Y}}(\widehat{\mu}(\lambda^*)) + \lambda^*(\mathcal{M}_{\bar{X}}(\widehat{\mu}(\lambda)) - v). \end{aligned}$$

Therefore, we have

$$h_X(v_1) - \lambda^*(v - v_1) \leq h_X(v)$$

for every v, which implies that $-\lambda^* \in \partial h_X(v_1)$. ■

APPENDIX D: THE DISCRETIZATION ALGORITHM (DA)

Consider the problem (11) for a fixed value of $u_1 = \bar{u}_1$. In this case, the computational problem reduces exactly to solving the EMEP with respect to X at fixed squared distance level $\bar{u}_1 - 1$. As shown in Section 3, such problem can be solved directly with the algorithm proposed in Appendix C.

Therefore, given the desired precision, one can discretize the range of the variable u_1, $[\bar{L}_1, \bar{U}_1]$, and solve the EMEP for each one of these values. Such scheme yields the following algorithm.

Discretization Algorithm
Input: Relative tolerance $\varepsilon > 0$, $u_1^1 = (1 + 2\varepsilon/N_1)\bar{L}_1$, $k = 1$.
Step 1. Evaluate $u_2^k = g(u_1^k)$ and compute $f^k = \frac{N_1}{2}\log(u_1^k) + \frac{N_2}{2}\log(u_2^k)$;
Step 2. If $(1 + 2\varepsilon/N_1)u_1^k > \bar{U}_1$, compute $f^{k+1} = \bar{U}_1\bar{L}_2$, goto Step 4;
Step 3. Else set $u_1^{k+1} \leftarrow (1 + 2\varepsilon/N_1)u_1^k$, $k \leftarrow k + 1$, goto Step 1;
Step 4. Report $\min_{1 \leq i \leq k} f^i$ and the correspondent pair $(\widehat{u}_1^{i*}, \widehat{u}_2^{i*})$.

The following complexity results hold for the Discretization Algorithm.

Theorem D.1 *The Discretization Algorithm reports an ε-solution for the original problem performing exactly $\lceil \log(\bar{U}_1/\bar{L}_1)/\log(1 + 2\varepsilon/N_1) \rceil$ loops.*

Proof. Let $u^* = (u_1^*, u_2^*)$ be a optimal solution. There exists a k such that $u_1^k < u_1^* < (1 + 2\varepsilon/N_1)u_1^k$. We consider f^{k+1} as our candidate. We have

$$\begin{aligned} f^* &= \tfrac{N_1}{2}\log(u_1^*) + \tfrac{N_2}{2}\log(u_2^*) \\ &\leq f^{k+1} = \tfrac{N_1}{2}\log(1 + 2\varepsilon/N_1) + \tfrac{N_1}{2}\log(u_1^k) + \tfrac{N_2}{2}\log(u_2^{k+1}) \qquad (26) \\ &\leq \varepsilon + \tfrac{N_1}{2}\log(u_1^k) + \tfrac{N_2}{2}\log(u_2^{k+1}) = \varepsilon + f^* \end{aligned}$$

where we also used that $u_2^{k+1} \leq u_2^*$, since g is decreasing.

The second statement of the Theorem follows by noting that we have $u_1^k = \bar{L}_1(1 + 2\varepsilon/N_1)^k \leq \bar{U}_1$ and by taking logs to bound k. ■

By choosing a sequence $\varepsilon_k \to 0$ we obtain a sequence of ε_k-solutions that converge to the optimal solution of the BFP. One drawback to this method is that it requires solving the EMEP at every point of the discretization. In practice, such requirement may be cumbersome.

APPENDIX E: MONTE CARLO STUDY OF SIZE

Small Size Instances			Size of the Test α											
			$\alpha = 0.10$				$\alpha = 0.05$				$\alpha = 0.01$			
d	N_1	N_2	W	LR	LM	B	W	LR	LM	B	W	LR	LM	B
2	10	20	0.160	0.133	0.102	0.092	0.106	0.077	0.046	0.046	0.039	0.020	0.005	0.009
2	20	40	0.138	0.122	0.106	0.103	0.081	0.067	0.050	0.051	0.023	0.013	0.007	0.009
2	40	80	0.120	0.114	0.110	0.107	0.069	0.061	0.055	0.054	0.017	0.014	0.011	0.011
5	25	50	0.171	0.133	0.098	0.090	0.101	0.073	0.047	0.046	0.035	0.019	0.005	0.008
5	50	100	0.124	0.110	0.094	0.090	0.068	0.055	0.041	0.041	0.017	0.011	0.007	0.008
5	100	200	0.113	0.106	0.098	0.096	0.063	0.057	0.053	0.052	0.015	0.013	0.010	0.010
10	50	100	0.175	0.131	0.094	0.084	0.102	0.072	0.044	0.041	0.035	0.018	0.008	0.008
10	100	200	0.137	0.118	0.099	0.093	0.074	0.062	0.047	0.047	0.019	0.012	0.009	0.009
10	200	400	0.116	0.107	0.100	0.098	0.062	0.056	0.051	0.051	0.014	0.011	0.009	0.009

Medium Size Instances			Size of the Test α											
			$\alpha = 0.10$				$\alpha = 0.05$				$\alpha = 0.01$			
d	N_1	N_2	W	LR	LM	B	W	LR	LM	B	W	LR	LM	B
25	125	250	0.196	0.141	0.096	0.077	0.118	0.079	0.043	0.034	0.035	0.018	0.007	0.007
25	250	500	0.137	0.109	0.088	0.078	0.071	0.056	0.039	0.036	0.019	0.014	0.008	0.007
25	500	1000	0.120	0.110	0.099	0.091	0.064	0.055	0.049	0.046	0.015	0.011	0.009	0.008
50	250	500	0.232	0.158	0.096	0.065	0.144	0.089	0.040	0.027	0.041	0.018	0.005	0.005
50	500	1000	0.147	0.117	0.091	0.079	0.083	0.061	0.044	0.038	0.021	0.013	0.008	0.007
50	1000	2000	0.126	0.111	0.100	0.092	0.070	0.062	0.053	0.049	0.016	0.012	0.011	0.010
75	375	750	0.262	0.170	0.098	0.064	0.167	0.098	0.048	0.029	0.059	0.025	0.008	0.004
75	750	1500	0.166	0.131	0.097	0.084	0.098	0.072	0.048	0.038	0.025	0.016	0.009	0.007
75	1500	3000	0.133	0.119	0.102	0.092	0.073	0.064	0.053	0.049	0.018	0.014	0.010	0.009
100	500	1000	0.284	0.175	0.090	0.054	0.179	0.097	0.043	0.025	0.060	0.025	0.007	0.004
100	1000	2000	0.175	0.134	0.101	0.076	0.104	0.071	0.047	0.036	0.026	0.016	0.008	0.006
100	2000	4000	0.139	0.117	0.099	0.087	0.073	0.061	0.050	0.042	0.017	0.013	0.009	0.007

Large Size Instances			Size of the Test α											
			$\alpha = 0.10$				$\alpha = 0.05$				$\alpha = 0.01$			
d	N_1	N_2	W	LR	LM	B	W	LR	LM	B	W	LR	LM	B
200	1000	2000	0.373	0.213	0.095	0.040	0.251	0.123	0.043	0.015	0.101	0.030	0.007	0.002
200	2000	4000	0.203	0.136	0.085	0.060	0.112	0.073	0.042	0.029	0.032	0.016	0.009	0.005
200	4000	8000	0.153	0.128	0.099	0.085	0.084	0.064	0.049	0.039	0.019	0.014	0.010	0.007

TABLE 2

Monte Carlo study of size for the Wald (W), Likelihood Ratio (LR), Lagrangian Multiplier (LM), and the Barlett Correction (B) of the Likelihood Ratio Tests (runs per entry = 10,000).

References.

[1] W. Behrens. Ein beitrag zur fehlerberechnung bei wenigen beobachtungen [a contribution to error estimation with few observations]. *Landwirtschaftliche Jahrbücher*, 68:807–837, 1929.

[2] E. Berndt and N. E. Savin. Conflict among criteria for testng hypothesis in the multivariate linear regression model. *Econometrica*, 45:1263–1278, 1977.

[3] J. Frédéric Bonnans, J. Charles Gilbert, Claude Lemaréchal, and Claudia A. Sagastizábal. *Numerical optimization*. Universitext. Springer-Verlag, Berlin, 2003. Theoretical and practical aspects, Translated and revised from the 1997 French original.

[4] T. S. Breusch. Conflict among criteria for testng hypothesis: extensions and comments. *Econometrica*, 47(1):203–208, 1979.

[5] M.-L. G. Buot and D. St. P. Richards. Counting and locating the solutions of polynomial systems of maximum likelihood equations, ii: The behrens-fisher problem. *submitted for publication*, 2006.

[6] A. R. Conn, N. I. M. Gould, and P. L. Toint. *Trust-region methods*. MPS/SIAM Series on Optimization. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2000.

[7] D. R. Cox. Effective degrees of freedom and the likelihood ratio test. *Biometrika*, 71(3):487–493, 1984.

[8] R. A. Fisher. The fiducial argument in statistical inference. *Annals of Eugenics*, 6:391–398, 1935.

[9] R. A. Fisher. The comparison of samples with possibly unequal variances. *Annals of Eugenics*, 9:174–180, 1939.

[10] L. G. Godfrey. *Misspecification tests in econometrics*. Cambridge University Press, New York, NY, 1988.

[11] W. H. Greene. *Econometric analysis*. Prentice-Hall Inc. Upper Saddle River, NJ, 3^{rd} edition, 1997.

[12] J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms*. Number 305-306 in Grund. der math. Wiss. (two volumes). Springer-Verlag, 1993.

[13] J. E. Kelley. The cutting plane method for solving convex programs. *J. Soc. Indust. Appl. Math.*, 8:703–712, 1960.

[14] S.-H. Kim and A. Cohen. On the behrens-fisher problem: a review. *Journal of Educational and Behavioral Statistics*, 23(4):356–377, 1998.

[15] S-J. Kim. A practical solution to the multivariate behrens-fisher problem. *Biometrika*, 79(1):171–176, 1992.

[16] K. Krishnamoorthy and J. Yu. Modified nel and van der merwe test for the multivariate behrens-fisher problem. *Statistics and Probability Letters*, 66:161–169, 2004.

[17] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Probability and Mathematical Statistics: A Series of Monographs and Textbooks. Academic Press, London-New York-Toront, Ont., 1979.

[18] Y. Nesterov and A. Nemirovskii. *Interior Point Polynomial Algorithms in Convex Programming*. Studies in Applied and Numerical Mathematics 13. Society for Industrial and Applied Mathematics (SIAM), 1994.

[19] H. Robbins, G. Simons, and N. Starr. A sequential analogue of the behrens-fisher problem. *The Annals of Mathematical Statistics*, 38(5):1384–1391, 1967.

[20] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.

[21] N. E. Savin. Conflict among testing procedures in a linear regression model with autoregressive disturbances. *Econometrica*, 44:1303–1315, 1976.

[22] H. Scheffé. On solutions of the behrens-fisher problem, based on the t-distribution. *The Annals of Mathematical Statistics*, 14(1):35–44, 1943.

[23] H. Scheffé. Practical solutions of the behrens-fisher problem. *Journal of the American Statistical Association*, 65(332):1501–1508, 1970.

[24] S. Smale. *Newton's method estimates from data at one point*. The Merging of Disciplines: New Directions in Pure, Applied and Computational Mathematics. 1986.

[25] A. Stuart and J. K. Ord. *Kendall's advanced theory of statistics*. Imprint London: Edward Arnold, 6^{th} edition, 1994.

[26] K. Subrahmaniam and K. Subrahmaniam. On the multivariate behrens-fisher problem. *Biometrika*, 60(1):107–111, 1973.

[27] B. L. Welch. The signficance of the difference between two means when the population variances are unequal. *Biometrika*, 29:350–362, 1938.

[28] B. L. Welch. The generalization of 'student's' problem when several different population variances are involved. *Biometrika*, 34:28–35, 1947.

[29] Y. Yanagihara and Ke-Hai Yuan. Three approximate solutions to the multivariate Behrens-Fisher problem. *Communications in Statistics - Simulation and Computation*, (34):975–988, 2005.

[30] Y. Yao. An approximate degrees of freedom solution to the multivariate behrens-fisher problem. *Biometrika*, 52(1/2):139–147, 1965.

[31] Y. Ye. *A new complexity result for minimizing a general quadratic function with a sphere constraint*. Recent advances in global optimization. 1992.

E-mail: belloni@mit.edu
1101 Kitchawan Road,
Building 800 Office 32-221,
Yorktown Heights, NY 10598

E-mail: didier@unc.edu
Smith Building CB#3260,
UNC-Chapel Hill
Chapel Hill, NC 27599